

# Data Set and Evaluation of Automated Construction of Financial Knowledge Graph

Wenguang Wang<sup>†</sup>, Yonglin Xu, Chunhui Du, Yunwen Chen, Yijie Wang & Hui Wen

DataGrand Inc., Shanghai 201203, China

**Keywords:** Knowledge graph; Entity extraction; Relation extraction; FR2KG data set; CCKS

Citation: Wang, W.G., et al.: Data set and evaluation of automated construction of financial knowledge graph. Data Intelligence 3(3), 418-443 (2021). doi: 10.1162/dint\_a\_00108

Received: February 11, 2021; Revised: March 20, 2021; Accepted: April 22, 2021

## ABSTRACT

With the technological development of entity extraction, relationship extraction, knowledge reasoning, and entity linking, the research on knowledge graph has been carried out in full swing in recent years. To better promote the development of knowledge graph, especially in the Chinese language and in the financial industry, we built a high-quality data set, named financial research report knowledge graph (FR2KG), and organized the automated construction of financial knowledge graph evaluation at the 2020 China Knowledge Graph and Semantic Computing Conference (CCKS2020). FR2KG consists of 17,799 entities, 26,798 relationship triples, and 1,328 attribute triples covering 10 entity types, 19 relationship types, and 6 attributes. Participants are required to develop a constructor that will automatically construct a financial knowledge graph based on the FR2KG. In addition, we summarized the technologies for automatically constructing knowledge graphs, and introduced the methods used by the winners and the results of this evaluation.

## 1. INTRODUCTION

With the advancement of technologies such as entity extraction, relation extraction, knowledge reasoning, and entity linking, the research into knowledge graph has been carried out in full swing in recent years. However, to date, there are relatively insufficient technologies for systematically and automatically constructing knowledge graphs. Moreover, the ability to automatically construct knowledge graphs determines the popularity of the application of knowledge graph technologies. To promote the further development of knowledge graph technologies, a good data set and an evaluation system are required. In

<sup>†</sup> Corresponding author: Wenguang Wang (Email: wangwenguang@datagrand.com; ORCID: 0000-0002-9617-0818).

this regard, ImageNet [1] is a classic example that has significantly promoted the development of computer vision. Similarly, in the field of knowledge graphs, the Text Analysis Conference (TAC) [2] released multiple data sets and organized corresponding evaluations to continuously promote the development of knowledge graphs. However, there is no similar data set or evaluation in the Chinese language.

One of the most widely used domain-specific knowledge graphs is financial knowledge graph, which is widely used in investment research, risk tracking and control, corporate public opinion management, intelligent question answering, and industrial analysis. In the financial field, constructing financial knowledge graphs from various unstructured text is a basic task of great value. However, there are insufficient studies [3, 4] that have constructed financial knowledge graphs. In particular, there is no data set for the automated construction of financial knowledge graphs and the corresponding evaluation.

To better promote the development of knowledge graphs, especially in Chinese and in the financial industry, we organized the automated construction of financial knowledge graph evaluation at the 2020 China Knowledge Graph and Semantic Computing Conference (CCKS2020). The data source for the evaluation was Chinese financial research reports on macroeconomics, industries, and companies conducted by professional financial institutions. The reports are characterized by comprehensiveness, high reliability, in-depth, and high quality. The scope of the content encompasses financial indicators, policies, and rich data, which are highly suitable for building a financial knowledge graph for financial institutions, governments, research institutes, etc., to provide an in-depth analysis and intelligent decision-making support. However, because the data and knowledge of financial research reports cover a wide range of areas and contain professional knowledge, different people express the same content differently, which makes it quite difficult to construct a knowledge graph from financial research reports. Solving these problems can greatly boost the development of automated construction of knowledge graphs, and have great academic value. We collected 1,200 financial research reports, and designed a knowledge graph schema with 10 entity types, 19 relationship types, and 6 attributes. In addition, we annotated 17,799 entities, 26,798 relationship triples <entity, relationship, entity>, and 1,328 attribute triples <entity, attribute key, attribute value> by experts in the financial industry based on the knowledge graph schema, which is the largest automated knowledge graph construction and evaluation data set that has been published in Chinese. Evaluation refers to the TAC 2016 Cold Start KBP Track plan [5], starting with a predefined knowledge graph schema and a seed knowledge graph, and automatically extracting entities, relationship triples, and attribute triples from the unstructured texts of financial research reports. The evaluation does not limit the algorithms and models used. Participants can use open pre-training models, such as BERT [6] and ERINE [7], and use third-party open knowledge graphs, such as those from OpenKG<sup>①</sup>. Simultaneously, the participants are encouraged to use various methods, such as unsupervised, weak supervision, and distant supervision, to realize the automated construction of knowledge graphs.

The remainder of this paper is organized as follows. Section 2 presents the evaluation tasks and methods with the release of a professional Financial Research Report Knowledge Graph (FR2KG) data set. The

<sup>①</sup> <http://www.openkg.cn/>

properties of the FR2KG data set are described in Section 3. Section 4 surveys the current knowledge graph construction technologies, and Section 5 introduces the methods used by the winners and the results of this evaluation. Finally, the challenges and prospects for the automated construction of domain knowledge graphs are discussed in Section 6.

## 2. TASK DEFINITION AND EVALUATION METRICS

### 2.1 Task Definition

The content of this evaluation is constructing a financial knowledge graph from the text of unstructured financial research reports, which is based on the given knowledge graph schema:

- Given: unstructured text of the financial research reports
- Given: the schema of knowledge graph
- Given: seed knowledge graph
- Participants are required to develop a constructor that extracts entities, attribute triples, and relationship triples that conform to the schema from the unstructured text provided.

There are 1,200 financial research reports. After removing tables, images, headers, footers, and other useless and repetitive information, the remaining contents are converted into plain text format. Simultaneously, we worked with financial research experts to analyze these research reports, and designed a schema of the financial knowledge graph based on the characteristics of financial research and technical evaluation. Next, the unstructured text was annotated by trained annotators, and the results were reviewed by financial research experts. Therefore, the annotated knowledge graph (annotated KG) is composed of a data set that has been reviewed. The annotated KG will be randomly divided into seed knowledge graph (seed KG) and evaluation knowledge graph (evaluation KG). The random segmentation method was as follows:

- 1) Randomly select 200 copies of 1,200 TXT files;
- 2) Select the extracted entities, relationship triples, and attribute triples, corresponding to 200 TXT files as seed KG; and
- 3) Remove all the data in the seed KG from the annotated KG, and use the remaining data as the evaluation KG.

The above is a complete description of the data processing and annotation process. Therefore, we have obtained the entire FR2KG data set, including knowledge graph schema, unstructured text financial research reports, seed KG, and evaluation KG. The goal of this evaluation is to use FR2KG to develop a financial knowledge graph constructor to automatically extract entities, attribute triples, and relationship triples from unstructured text. The constructed financial knowledge graph excludes the data that already exist in the seed KG, and the evaluation procedure uses the metrics described in the next section to measure the quality of the knowledge graph. The entire process mentioned above is shown in Figure 1.

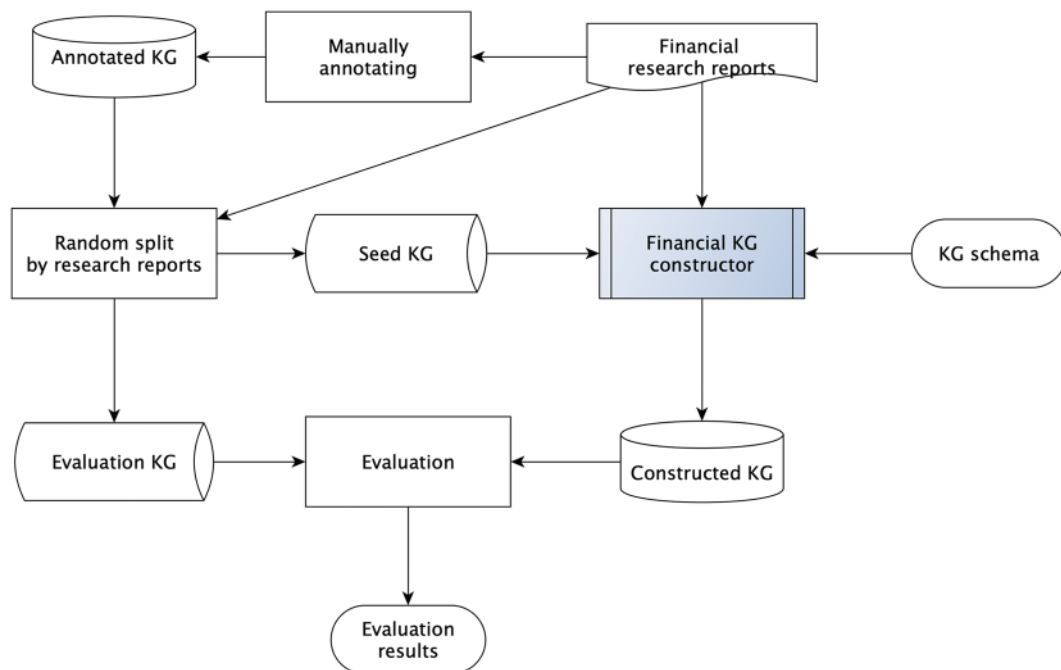


Figure 1. Schematic of the task process.

Given the FR2KG data set, the goal of the participants is to develop a financial knowledge graph constructor that is the most efficient at automatically extracting entities, attribute triples, and relationship triples from unstructured text of financial research reports, and constructing a financial knowledge graph that is as consistent as possible with the knowledge graph annotated by experts. To be as close to the real application scenario as possible, and considering the fairness and reasonableness of all participants, this evaluation allows all participants to use various open or public data, including, but not limited to, pre-trained models, open knowledge graphs from OpenKG, and other sources. If participants would like to use private data, the data must be publicly available, and other participants should be able to use them.

## 2.2 Evaluation Metrics

This evaluation task uses the  $F1$  score defined as follows to evaluate the performance of the knowledge graph constructor. The higher the  $F1$  score, the better is the performance. The data of the knowledge graph are divided into three types: entity, attribute triples  $\langle \text{entity}, \text{attribute key}, \text{attribute value} \rangle$ , and relationship triples  $\langle \text{entity}, \text{relationship}, \text{entity} \rangle$ . Precision ( $p$ ), recall ( $r$ ), and  $F1$  score ( $F1$ ) are defined as follows. First, we define the following variables:

$y_E$ : The set of all the pairs of  $\langle \text{entity}, \text{entity type} \rangle$  extracted by the constructor;  $|y_E|$  represents the number of entities.

$\hat{y}_E$ : The set of all pairs of <entity, entity type> annotated by experts;  $|\hat{y}_E|$  represents the number of entities.

$y_A$ : The set of all attribute triples <entity, attribute key, attribute value> extracted by the constructor;  $|y_A|$  represents the number of triples.

$\hat{y}_A$ : The set of all attribute triples <entity, attribute, attribute type> annotated by experts;  $|\hat{y}_A|$  represents the number of triples.

$y_R$ : The set of all relation triples <entity, relationship, entity> extracted by the constructor;  $|y_R|$  represents the number of triples.

$\hat{y}_R$ : The set of all relation triples <entity, relation, entity> annotated by the expert;  $|\hat{y}_R|$  represents the number of triples.

$y \cap \hat{y}$ : The intersection of  $y$  and  $\hat{y}$ , that is, the same part extracted by the constructor as the expert annotated, which represents the entities, attribute triples, or relationship triples correctly extracted by the constructor.  $y \cap \hat{y}$  represents the number.

Thus, we have the following:

Entities:

$$p_E = \frac{|y_E \cap \hat{y}_E|}{|\hat{y}_E|} \quad (1)$$

$$r_E = \frac{|y_E \cap \hat{y}_E|}{|\hat{y}_E|} \quad (2)$$

$$F1_E = 2 \times \frac{p_E \times r_E}{p_E + r_E} \quad (3)$$

Attribute triples:

$$p_A = \frac{|y_A \cap \hat{y}_A|}{|\hat{y}_A|} \quad (4)$$

$$r_A = \frac{|y_A \cap \hat{y}_A|}{|\hat{y}_A|} \quad (5)$$

$$F1_A = 2 \times \frac{p_A \times r_A}{p_A + r_A} \quad (6)$$

Relationship triples:

$$p_R = \frac{|y_R \cap \hat{y}_R|}{|y_R|} \quad (7)$$

$$r_R = \frac{|y_R \cap \hat{y}_R|}{|\hat{y}_R|} \quad (8)$$

$$F1_R = 2 \times \frac{p_R \times r_R}{p_R + r_R} \quad (9)$$

Finally, we define the final evaluation  $F1$ -score of the entire knowledge graph as in Equation (10):

$$F1 = \frac{F1_E + 2 \times F1_A + 2 \times F1_R}{5} \quad (10)$$

We use the weighted average of the  $F1$ -score of three types of entity, attribute triple, and relationship triple, in which the weights from attribute triples and relation triples are twice that of the entities, because we believe extracting attribute triples and relation triples is twice as difficult as extracting entities. For an attribute triple, both the entity and attribute value must be extracted correctly, and the attribute value should match the attribute key, which is the same as identifying the entity and matching the entity type. So, we decide that the weight of attribute extraction is twice that of entity extraction. For a relation triple, it is necessary to extract two entities correctly and match the corresponding relationship type. It is a significant topic to study the difficulty of entity extraction, attribute extraction and relationship extraction in detail. In determining the evaluation metrics, we carried out a survey and did not find the corresponding research. So, we decided the weight of 2 based on our experience.

### 3. PROPERTIES OF FR2KG

#### 3.1 FR2KG Overview

Among data sets that have been constructed, some are manually annotated [8, 9], some are collaboratively annotated by humans and algorithms and others are labeled with higher precision through better algorithms. However, most of the data sets are concentrated, in general, with news and common-sense articles (such as Wikipedia), as well as in some domains, such as biomedical and medical-related and scientific and technological literature data sets. Data sets for financial knowledge graphs are rare, and Chinese financial knowledge graph data sets are even rarer. For the first time in this evaluation, the FR2KG data set and the corresponding unstructured texts are published, aiming to promote the development of technologies for distant supervision or weak supervision, to automatically construct domain knowledge graphs.

The construction process of the FR2KG data set is described in the previous section, as shown in Figure 1. First, 1,200 financial research reports were collected. Experts in the financial field analyzed these reports, extracted the plain text from the main body, and saved it in the TXT format as the basic unstructured

text corpus. Then, the experts and the knowledge graph team studied these corpora together, designed the schema of the knowledge graph from the perspective of financial business, and performed iterative optimization according to the characteristics of the evaluation, and finally, determined that it contained 10 entity types, 6 entity attributes, and 19 relationships between the entities. Subsequently, these corpora were annotated with the help of the annotation system of the Yuanhai Knowledge Graph Platform, which is a product of DataGrand Inc. The annotation system is specifically used for the annotation of the knowledge graph, and supports the annotation of entities, entity attributes, and relationships between entities. Before annotating, all annotators were trained by financial experts to align their understanding of the schema. All annotated data were reviewed by experts, and then, divided into seed KG and evaluation KG, as described in the previous section. Examples of the FR2KG data set are shown in Figures 2 and 3.

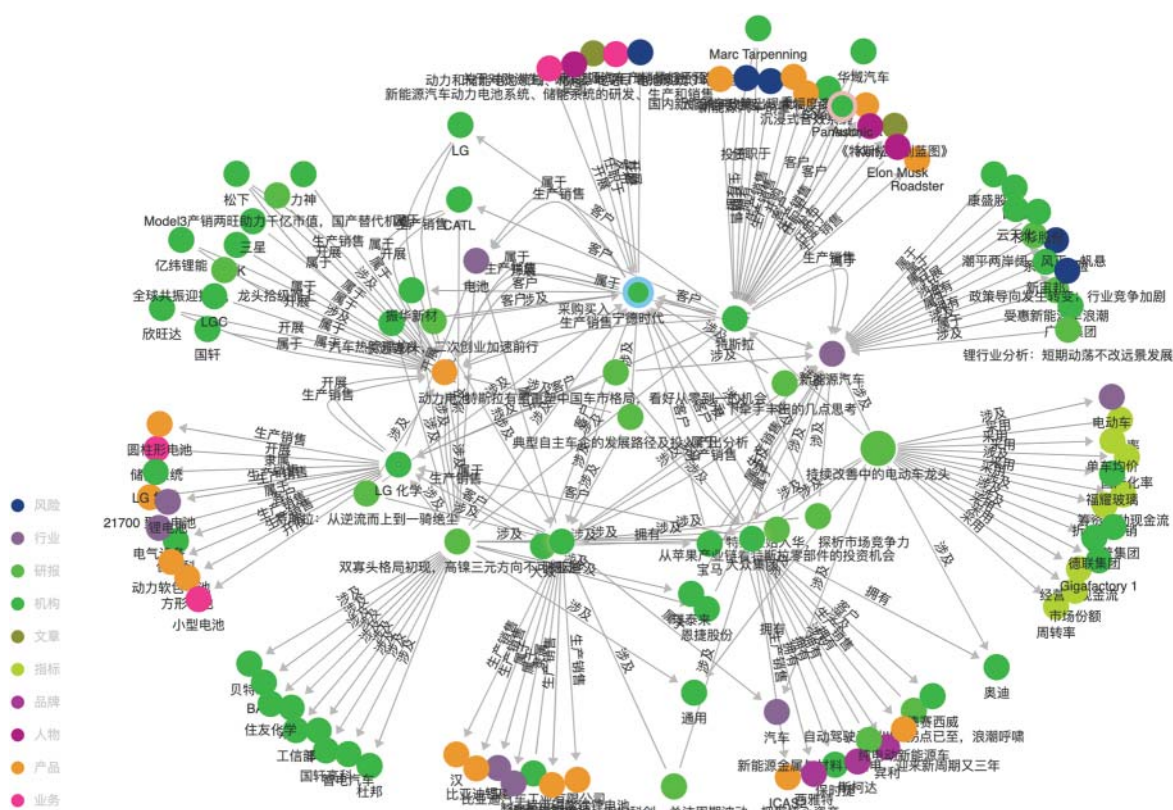


Figure 2. An example of the FR2KG in Yuanhai Knowledge Graph Platform.

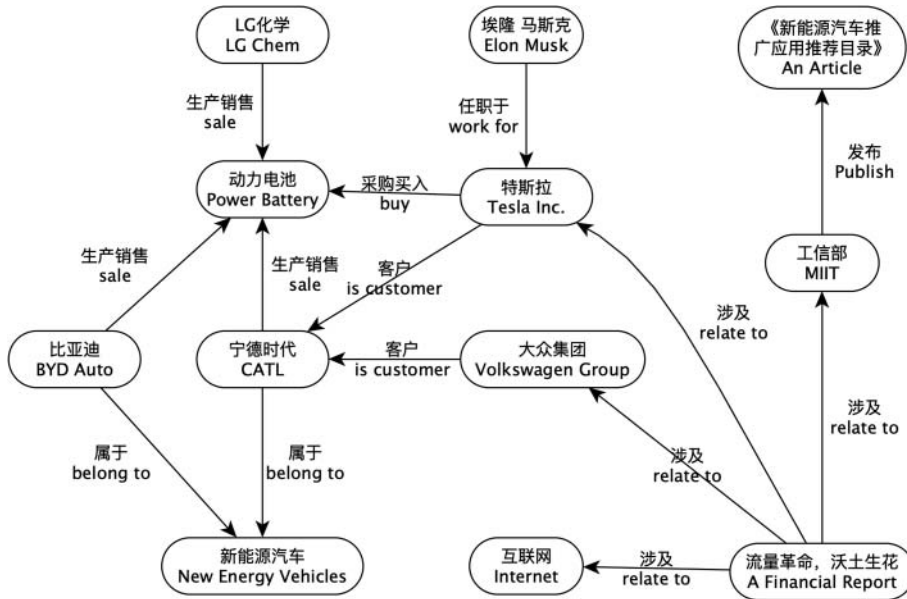


Figure 3. A simplified example of FR2KG.

A summary of FR2KG is shown in Table 1. It is currently the largest data set for the automatic construction of Chinese financial knowledge graphs. Table 1 describes the data, and the following sections introduce FR2KG in detail.

Table 1. Summarization of FR2KG.

	Entities number	Relationship triples number	Attribute triples number
Seed KG	5,131	6,091	354
Evaluation KG	12,668	20,707	974

### 3.2 Financial Research Reports

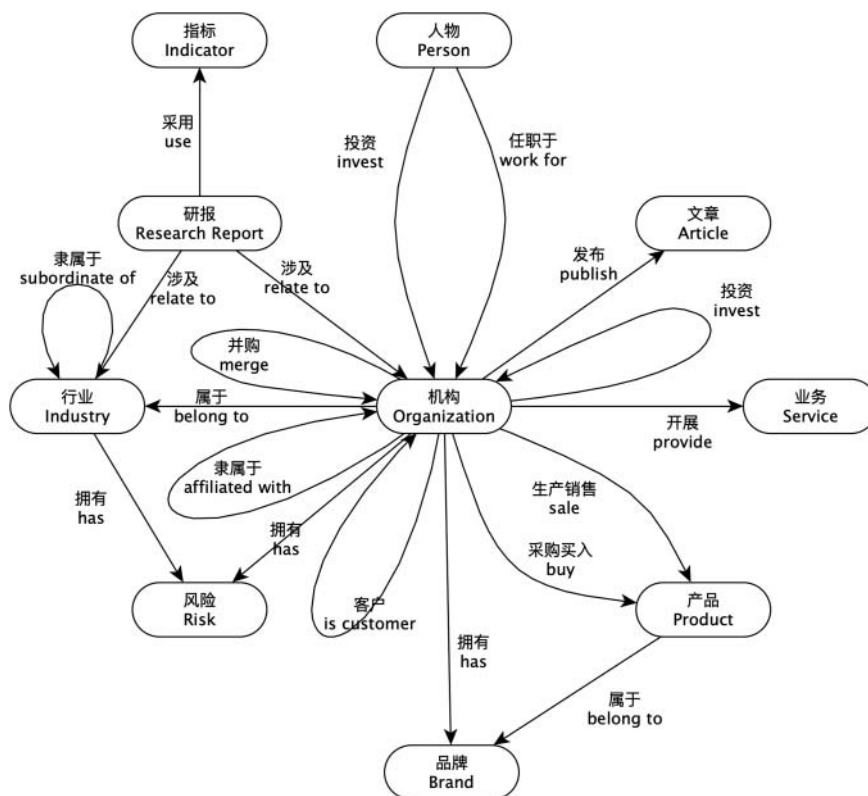
The length of the financial research reports varies. As shown in Table 2, the longest text has 13,857 characters, while the shortest has only 242 characters, and the longest is close to 60 times the shortest. However, the length of most texts is concentrated in the range of 1,000–3,000 fields, accounting for 70% of the total text. In terms of paragraphs, the shortest has 4 paragraphs, and the longest has 74 paragraphs. Most reports were between 10 and 30 paragraphs, accounting for 82% of the total text.

**Table 2.** Statistics of 1,200 financial research report texts in FR2KG.

	By characters	By paragraphs
Mean	2,214	19.9
Standard Deviation	1,173	9.5
Minimum	242	4
First Quartile	1,392	14
Median	2,018	18
Third Quartile	2,779	24
Maximum	13,857	74

### 3.3 Schema

Figure 4 shows the schema of FR2KG. There are 10 entity types in total, which are represented by ellipses, and 19 relationships between entity types, which are represented by directed arrows. For example, in the relationship of <人物/person, 投资/investment, 机构/organization>, the directed arrow in the figure points from "person" to "organization". Among these entity types, the three entity types have attributes (Table 3). Notably, the attribute value of the time type is normalized to the "YYYY-mm-dd" format during annotation. The participants were also required to normalize the time data in the construction of the knowledge graph.

**Figure 4.** The knowledge graph schema of FR2KG.

**Table 3.** Attributes of entity types of FR2KG schema.

Entity type	Attribute key	Data type of attribute value
研报/Research Report	发布时间/Publish Date	Date
	评级/Rating	String
	上次评级/Previous Rating	String
机构/Organization	全称/Full Name	String
	英文名/English Name	String
文章/Article	发布时间/Publish Date	Date

The FR2KG schema, as shown in Figure 4, is very rich in applications that can be used in investment research, financial risk assessment and control, product analysis, industrial chain analysis, and other fields. For example, with the relationships of <人物/person, 投资/invest, 机构/organization> and <机构/organization, 投资/invest, 机构/organization>, it can be used for in-depth investment and financing analysis. Another example is the relationship between <机构/organization, 生产销售/sale, 产品/product> and <机构/organization, 采购买入/buy, 产品/product>, which can be used for supply chain analysis, mining the company's advantages in the supply chain and assessing supply chain risks.

### 3.4 Entities and Attributes

Tables 4 and 5 summarize the entities and their attribute triples of FR2KG.

**Table 4.** Statistics of entities of FR2KG.

Entity Type	Number of seed KG	Number of evaluation KG
人物/Person	86	383
行业/Industry	552	1,253
业务/Service	664	1,053
产品/Product	1,218	3,366
研报/Research Report	73	307
机构/Organization	1,742	3,739
风险/Risk	243	826
文章/Article	170	450
指标/Indicator	239	856
品牌/Brand	144	435
	51,31	12,668

**Table 5.** Statistics of attribute triples of FR2KG.

Entity type	Attribute key	Number of seed KG	Number of evaluation KG
研报/Research Report	发布时间/Publish Date	70	284
	评级/Rating	63	216
	上次评级/Previous Rating	46	155
机构/Organization	全称/Full Name	90	35
	英文名/English Name	31	111
文章/Article	发布时间/Publish Date	54	173
		354	974

### 3.5 Relationships

Table 6 summarizes the statistics of relationship triples in the FR2KG.

**Table 6.** Statistics of relationship triples of FR2KG.

Head entity type	Relationship	Tail entity type	Number of triples of seed KG	Number of triples of evaluation KG
产品/Product	采用/Use	品牌/Brand	100	189
机构/Organization	采用/Use	行业/Industry	540	2,081
机构/Organization	投资/Invest	机构/Organization	88	239
机构/Organization	拥有/Has	品牌/Brand	85	331
机构/Organization	隶属于/Affiliated with	机构/Organization	120	401
行业/Industry	隶属于/Subordinate of	行业/Industry	109	485
机构/Organization	客户/Is customer	机构/Organization	159	243
机构/Organization	并购/Merge	机构/Organization	65	196
机构/Organization	发布/Publish	文章/Article	111	287
人物/Person	投资/Invest	机构/Organization	7	51
人物/Person	任职于/Work for	机构/Organization	54	259
机构/Organization	开展/Provide	业务/Service	563	1,235
机构/Organization	采购/Buy	产品/Product	45	75
机构/Organization	生产销售/Sale	产品/Product	758	2,232
机构/Organization	拥有/Has	风险/Risk	171	422
行业/Industry	拥有/Has	风险/Risk	68	431
研报/Research Report	采用/Use	指标/Indicator	411	1,736
研报/Research Report	涉及/Relate to	行业/Industry	748	1,658
研报/Research Report	涉及/Relate to	机构/Organization	1,889	8,156
			6,091	20,707

### 3.6 Properties of FR2KG

As a complete domain knowledge graph data set, FR2KG is currently the largest Chinese financial knowledge graph data set dedicated to the automated construction of knowledge graphs. In the future, we plan to continue expanding and enriching its content.

**Scale:** FR2KG is committed to promoting the development of automated construction of domain knowledge graphs, including rich and diverse data types and the largest data scale currently. In addition, the content is abundant, including common stock research reports, industry research reports, and macroeconomic research reports in the financial industry.

**Relationship:** The data set provides 19 common relationships in the financial field, as shown in Figure 4. These relationships can help realize multiple and diverse analyses in the financial industry. For example, through the relationship of <机构/organization, 拥有/has, 风险/risk> and <行业/industry, 拥有/has, 风险/risk>, industry or enterprise risk analysis, risk assessment, and risk early warning could be performed better. The relationship between <人物/person, 任职于/work for, 机构/organization>, <人物/person, 投资/invest,

机构/organization>, and <机构/organization, 投资/invest, 机构/organization> can apply to deep-level equity relationship mining, which has great value in bank loans and investment analysis.

**Professionalism:** The schema of FR2KG is jointly designed by experts in the financial industry and knowledge graph experts. The annotators were trained by financial experts before annotation, and the results were reviewed by financial experts to ensure the high professionalism of the data set.

**Diversity:** The goal of FR2KG is to evaluate the performance of the automated construction of financial knowledge graphs; however, the application of the data set is not limited to this objective. Various other technologies related to the knowledge graphs can also be evaluated using FR2KG. For example, common tasks, such as link prediction and node classification in graph neural networks, tasks related to various graph algorithms, and tasks based on deep learning techniques to implement traditional graph algorithms, can be evaluated.

## 4. RECENT TRENDS IN KNOWLEDGE GRAPH AUTOMATION CONSTRUCTION

### 4.1 Entity Extraction

Entity extraction, also known as named entity recognition (NER), aims to recognize the mentions of rigid designators from text belonging to predefined semantic types such as person, location, and organization. The two popular data sets from recent work, CoNLL03 [14] and OntoNotes5.0 CoNLL03 contain annotations for Reuters' news in English and German. The English data set contains a large portion of sports news with annotations in four entity types: person, location, organization, and miscellaneous entities. OntoNotes5.0 contains annotations for a large corpus, comprising various genres with structural information and shallow semantics. The data set was annotated using 18 entity types. BOSON [15], People's Daily [16], and MSRA [17] are Chinese entity extraction data sets in general fields, while FR2KG proposed in this article focuses on the Chinese financial field.

In recent years, research on supervised entity extraction has mainly been focused on how to input representation and design neural models, including context encoders and tag decoders. In addition, unsupervised and semi-supervised entity extraction has achieved remarkable development.

#### 4.1.1 Supervised Entity Extraction

Input representation is the first step in the entity extraction. In this subsection, we summarize word-level representation, character-level representation, language model, and other representations. Since Mikolov et al. [18] proposed word2vec, many studies on entity extraction have used the word2vec toolkit to train word-level representation on different corpora, such as PubMed [19], Gigaword [20], NYT [21], and SENNA [22]. In addition, GloVe [23] and FastText [24] are widely used. Instead of only considering word-level representations, character-level representation has been found to be useful for exploiting explicit sub-word-level information and naturally handling out-of-vocabulary information [21, 25, 26]. Both word-level and character-level representations only contain the meaning of the word, without its context.

Therefore, many studies have added context-dependent language model representations to the input representation. Peters et al. [27] proposed TagLM, a language model augmented sequence tagger. This tagger considers both pre-trained word embeddings and bidirectional language model embeddings for each token in the input sequence. Based on TagLM, Peters et al. [26] proposed the famous pre-trained bidirectional language model ELMo. The key difference between ELMo and TagLM is that ELMo allows the task model to learn a weighted average of all bidirectional LM layers, whereas TagLM only uses the top bidirectional LM layer. In contrast to CNNs and recurrent neural networks (RNNs), transformers [28] utilize stacked self-attention and point-wise, fully connected layers to build basic blocks for the encoder and decoder. Based on the transformer, BERT [6] was proposed to pre-train a deep bidirectional transformer by jointly conditioning both the left and right contexts in all layers. Combining pre-trained language model embedding with traditional embedding has become a de facto standard [29, 30, 31, 32, 33]. In addition, novel input representations are still being explored, such as external knowledge from Wikidata [30], dependency trees [31], and global contextual embedding [32, 33].

After converting the input sentence into a representation, the context encoder captures the context dependencies, and the tag decoder predicts tags for tokens in the input sequence. Collobert et al. [34] used a CNN to produce local features around each word, and applied a maximum or averaging operation to extract global features. Strubell et al. [35] proposed an iterated dilated CNN (ID-CNN), where four stacked dilated convolutions having a width of three obtained more contextual information. Compared to CNN, the bidirectional RNN makes full use of the forward and backward information in the sentence, which can effectively extract the features of the entire sentence. Therefore, a bidirectional RNN is the most popular encoder for entity extraction tasks. Although we can directly use the hidden layer of the bidirectional RNN to connect to the softmax layer, adding the CRF layer as the tag decoder can help in understanding the limitations of the sentence, to ensure the effectiveness of the prediction. Huang et al. [22] were the first to utilize the BiLSTM-CRF architecture to implement sequence tagging tasks, including POS, chunking, and NER. Similar to [22], many studies also used BiLSTM as an encoder and CRF as a decoder [21, 36, 37]. In addition to CRF, RNN [38] and pointer network (PtrNet) [39] have also been explored as tag decoders. Shen et al. [40] reported that RNN tag decoders outperform CRF and are faster to train when the number of entity types is large. However, a major disadvantage of RNN and PtrNet decoders lies in greedy decoding, meaning that the input of the current step requires the output of the previous step. Because the pre-trained model can capture sufficient semantic information, some studies only use BERT and completely abandon BiLSTM-CRF. In particular, Li et al. [41] framed the NER task as a machine reading comprehension (MRC) problem, which can be solved by fine-tuning the BERT model.

#### 4.1.2 Semi-supervised Entity Extraction

The semi-supervised entity extraction method aims to manually add a small number of appropriate entities as training corpus according to the entity type designed in advance by humans, and use the pattern learning method for continuous iterative learning and manual adjustments to finally generate a named entity data set, which reduces the dependence on manual annotation corpus. Liu et al. [42] used a small amount of existing labeled data to train the initial KNN and CRF models, performed semi-supervised learning on

tweet data, and improved the training data by learning and supplementing data with a large amount of unlabeled text. Etzioni et al. [43] proposed the KNOWITALL system, based on a set of predicate inputs, using pattern learning, subclass extraction, list extraction (uppercase), and other modes to perform NER on unlabeled data. In addition, Zhang and Elhadad [44] proposed a method for extracting named entities from biomedical text based on terminology, corpus statistics (such as inverse document frequency and context vector), and shallow syntactic knowledge (such as noun phrase chunks), and conducted experiments on two mainstream biomedical data sets to verify the method.

#### 4.1.3 Unsupervised Entity Extraction Methods

Based on the vocabulary resources, vocabulary patterns, and statistical data that are calculated on a large corpus, named entities can be inferred by clustering and combining the similarity in the sentence context. Nadeau et al. [45] proposed an unsupervised system for the construction of geographical name dictionaries and the resolution of named entity ambiguities. Zhang and Elhadad [44] proposed an unsupervised method that uses terminology, corpus statistics and shallow grammatical knowledge to extract named entities from biomedical texts, and proved the effectiveness and versatility of the unsupervised method. Brooke et al. [46] performed Brown clustering based on pre-segmented expectations, combined with the rank value of each class after clustering, and constructed bootstrap seeds for training, which can extract entities for specific domain knowledge. Jia et al. [47] used cross-domain language modeling and obtained task and domain vectors to complete NER entity extraction in unsupervised and supervised fields, respectively. Collins and Singer [48] only used seven simple “seed” rules to realize NER on the original data, and proposed two unsupervised named entity classification algorithms.

## 4.2 Relation Extraction

Relation extraction is usually considered to be a classification task, which predicts semantic relationships between pairs of nominals and can be defined as follows. Given sentence  $S$  with annotated pairs of nominals  $e_1$  and  $e_2$ , we aim to identify the relationships between  $e_1$  and  $e_2$ . Relation extraction is usually divided into supervised, unsupervised, and distant supervision relation extraction. End-to-end entities and relation extraction are also popular. Supervised data sets are of high quality and contain almost no noise, but are often small. SemEval2010 Task 8 [49] contained nine directed relation types and 10,717 samples, of which 8,000 were used for training and 2,717 for testing. ACE2005 contains 599 documents, which are related to news and e-mail and divided into seven main types of relations. Each type of relationship has an average of 700 instances for training and testing. In addition to ACE2005, which contains the Chinese corpus, DuIE [50] is another large-scale Chinese data set for information extraction. The FR2KG proposed in this study focuses on relation extraction in the Chinese financial field. For distant supervision relation extraction, the New York Time (NYT) data set is formed by aligning the relation with Freebase. The data set contains 52 possible relationship categories and a special relationship category NA (indicating that there is no relation between entities). The training data contain 522,611 sentences, 281,270 entity pairs, and 18,252 relations.

#### 4.2.1 Supervised Relation Extraction

Zeng et al. [51] used a CNN to extract vocabulary and sentence-level features for relation extraction tasks. The lexical-level feature vector is concatenated by the word vector of the labeled entity as well as the context and semantic category feature in WordNet. The sentence-level feature representation was automatically extracted using the maximum pooling CNN. In order to eliminate the impact of artificial class, Santos et al. [52] used a pairwise ranking loss function for training instead of cross entropy. Because there is a lot of irrelevant information in the sentence, the method of extracting sequence features cannot accurately predict the relationship between the two entities. Therefore, Xu et al. [53] noted that the shortest dependency path (SDP) is beneficial for determining the relationship between two entities. Specifically, [53] successively took the SDP from the subject to the object as input, passed it through the lookup table layer, produced local features around each node on the dependency path, and combined these features into a global feature vector through a CNN that was then fed to a softmax classifier. Similarly, Xu et al. [54] used a four-channel LSTM to extract words, parts of speech, grammatical relations, and WordNet semantic features along with the SDP. However, previous studies based on SDP may neglect crucial information. Zhang et al. [55] encoded a complete dependency structure over an input sentence with an efficient graph convolutional network (GCN), and then, extracted entity-centric representations to make robust relation predictions. To avoid the introduction of irrelevant information between entities in the complete dependency tree, Guo et al. [56] proposed AGGCN to automatically generate the substructures for relation extraction tasks.

#### 4.2.2 Distant Supervision Relation Extraction

Supervised relationship extraction requires a large amount of expert-labeled data, which limits the application of this method. Therefore, Mintz et al. [57] proposed the hypothesis of remote supervision as follows. If two entities have a relationship in a known knowledge base, then all sentences that mention these two entities will express that relationship in some way. They then applied this assumption to align the document with the existing database and automatically generate a large amount of training data. Zeng et al. [58] proposed a piecewise convolutional neural network (PCNN) to extract features, and used a multi-instance learning method to alleviate the data noise problem. However, they failed to fully utilize the information across different sentences, and ignored the fact that there can be multiple relationships between the same entity pair. Therefore, after using PCNN to extract the features of each sentence in the package, Jiang et al. [59] used cross-sentence maximum pooling to select the features of different sentences, and then, aggregated the most important features into the representation of each entity pair. Finally, the feature is applied as a sigmoid instead of softmax to judge the possibility of multiple labels. Since different sentences have different contributions, [60, 61] focused on how to use the attention mechanism to select sentences. Inspired by the transE model [62], for the two entities  $e_1$  and  $e_2$  of each package,  $e_1 - e_2$  is used to represent the relation between the two entities. The features extracted by PCNN and relation  $e_1 - e_2$  are concatenated to obtain the weight of each sentence, and the feature of the package is the weighted sum of all sentence feature vectors. Du et al. [63] proposed a new multi-layer structured self-attention model based on BiLSTM.

Among them, the word-level attention mechanism based on a two-dimensional matrix can focus on different aspects of a sentence to better learn the contextual representation. The two-dimensional sentence-level attention mechanism used for multi-example learning can focus on different effective examples to better select sentences. Many studies use existing knowledge bases to add information to alleviate the problem of mislabeling in remote supervision. Vashishth et al. [64] added an additional entity type and relationship as information from the knowledge base (KB) to improve prediction performance. In addition, Wang et al. [65] proposed a label-free distant supervision method that does not use the relation labels under this inadequate assumption, but only uses the prior knowledge derived from the KB to supervise the learning of the classifier directly and softly.

#### 4.2.3 Unsupervised Relation Extraction

The unsupervised learning method assumes that the entity pairs with the same semantic relationship have similar context information, and the corresponding context information of each entity pair can be used to represent the semantic relationship of the entity pair. Hasegawa et al. [66] clustered entity pairs with the same contextual semantics, and then selected a core vocabulary to mark the semantic relationship between the categories. [67] improved Hasegawa's hypothesis by eliminating candidate entity pairs with multiple relationships or performing multi-level clustering to extract relationships. Davidov et al. [68] used Google search as the knowledge background to define concept words; however, without pre-defining any relationship types in advance, they can automatically extract related entities and semantic relationships. Yan et al. [69] combined dependency features and shallow grammatical templates, and used clustering methods to extract all the semantic relationships of entities in Wikipedia entries from a large-scale corpus. In addition, Bollegala et al. [70] analyzed the templates after clustering, found the implicit semantic relationship between the entity pairs, and selected suitable extraction templates from the candidate relationship templates, which expanded the scope of entity relationships and improved the accuracy and recall rate to a certain extent.

#### 4.2.4 End-to-end Entity and Relation Extraction

Entity relation extraction can be pipelined for NER and relation classification. This independent framework is more flexible, but ignores the correlation between the two tasks. The result of entity recognition may affect the performance of the relationship classification and lead to incorrect transmission. In contrast to the pipeline method, the joint learning framework can use a single model to extract entities and relationships, and can effectively integrate information regarding entities and relations.

An end-to-end approach is to share the model parameters between the entity recognition task and relationship classification task. Miwa et al. [71] proposed an end-to-end tree model that captures both word sequence and dependency tree substructure information by stacking tree-structured LSTM on BiLSTM. Zheng et al. [21] designed novel tags that contain information regarding entities and the relationships they hold. Based on this tagging scheme, the joint extraction of entities and relations can be transformed into a tagging problem. However, it is difficult to solve the problem of overlapping triples of different relations in

sentences. Zeng et al. [72] divided the sentences into three types according to the triplet overlap degree: Normal, EntityPairOverlap, and SingleEntiyOverlap. They proposed an end-to-end model based on sequence-to-sequence learning using a copy mechanism. The encoder converts a natural language sentence into a fixed-length semantic vector, and then the decoder reads in this vector and generates multiple triplets. To better consider the interaction of different relations in sentences, especially overlapping relations, Fu et al. [73] proposed an end-to-end model, GraphRel. In the first stage, GraphRel learns to automatically extract hidden features for each word by stacking a BiLSTM sentence encoder and a GCN dependency tree encoder to tag entity mention words and predict relation triplets. In the second stage, GraphRel uses a novel relation-weighted GCN to better predict the interaction between triples.

The advantage of the shared model parameters is that there is no need to attach constraints to the two subtasks; however, the independent sub-models do not allow the relationship between the two subtasks to be fully utilized. Therefore, studies must be conducted for achieving global optimization of joint extraction. Based on sharing model parameters, Sun et al. [74] proposed a global loss function to explore the mutual influence of the entity and relational models. Most existing methods determine relation types only after all entities have been recognized so the interaction between relation types and entity mentions is not fully modeled. Takanobu et al. [75] applied a hierarchical reinforcement learning framework to enhance the interaction between entity mentions and relation types. The high-level process detects the relationship indicator at a specific location. If a relationship is determined, a low-level process is triggered to identify the entity corresponding to the relationship. When low-level tasks are completed, the high-level reinforcement learning process continues to search for the next relationship in the sentence. Li et al. [76] transformed the entity relationship extraction task into multiple rounds of questions and answers; that is, the entity and relationship extraction is transformed into a task of determining the answer from the context. This method provides a better way to capture the label hierarchy dependency. However, this intermediate method is computationally inefficient because it needs to scan all entity template questions and related relationship template questions in a single sentence.

## 5. PARTICIPANTS OVERVIEW AND EVALUATION RESULTS

A total of 740 teams participated in the evaluation. In the top 18 teams, three teams were from companies, 10 teams were from universities, three teams were a combination of universities and companies, and the other two teams did not disclose relevant information. Table 7 presents a summary of prize-winning teams.

**Table 7.** Summary of top five teams.

Rank	Team	Affiliation	F1 Score
1	UPSIDE-DOWN	State Grid Information & Telecommunication Group Co., Ltd.	0.49704
2	Solaris99	Peking University	0.48340
3	BOOMBOOM	Shanghai Jiao Tong University Beijing Yuannian Technology Co., Ltd. The University of Edinburgh	0.46455
4	SGIT	Fujian YIRONG Information Technology Co., Ltd.	0.45376
5	Iceburg	Beijing University of Posts and Telecommunications	0.41169

The top five teams in this evaluation have sorted out and submitted a brief description of the methods they used. These methods and descriptions are analyzed and summarized below.

- All teams used rule-based methods or labeling functions to produce a training corpus. Only one team manually labeled 20 research reports as supplementary and validation samples, in addition to the automatically generated samples.
- All teams used BERT-based models in entity extraction; in addition to models, rule-based methods are used to supplement specific entity types. One team used the BERT-softmax model, three teams used the BERT-CRF model architecture, and the other team used the BERT-MRC [38] architecture.
- In terms of relationship and attribute extraction, all teams used a method based on co-occurrence. Co-occurrence is the basic assumption of distant supervision; that is, when two entities appear together in a short text, it can be assumed that they have a corresponding relationship. Based on the assumption of co-occurrence, the three teams used rule-based methods to determine whether this relationship existed, and the other two teams used BERT-based models to classify the relationships.
- A team used a clustering method to cluster research reports on similar or the same topics.

Summarizing the methods used by these teams, in the tasks of entity, relationship, and attribute extraction of knowledge graphs, the method based on the BERT pre-training model is still the best and most popular currently; it is also widely used. Because this evaluation is very close to the real application scenarios of the industry, in addition to using the BERT-based model, the rule-based method is still very effective in some cases and is an effective complement to the algorithms.

## 6. CHALLENGES AND LOOKING AHEAD

From the results of this evaluation, the highest *F1* value is approximately 0.5, when automatically constructing a financial knowledge graph based on the predefined schema, which is far from the requirements of real applications. This sets out some more challenging topics and new directions for research in knowledge graphs.

- In the field of automatically constructing knowledge graphs with a given schema and seed knowledge graph, the existing methods are not very effective. Developing end-to-end methods or multi-step frameworks to automate the construction of knowledge graphs is still a difficult task.
- Given a knowledge graph with schema, it is worthwhile to determine how to automatically annotate training data for entity, attribute, and relationship extraction. In addition, with the automatically annotated training data, building an excellent model to construct a high-precision and high-recall knowledge graph is still a challenge.
- In terms of entity extraction, the prize-winning participants in the evaluation used the BERT-based model with the rule-based method. Further research should be conducted about the end-to-end model and unified framework for this real scenario.
- Relationship and attribute extraction is currently focused on the use of co-occurrence with a rule-based or model-based filter, which highly depends on the performance of entity extraction. Entity extraction, having high precision and high recall, results in good relationships and attributes extraction. However, a method for achieving a good relationship and attribute extraction when there is considerable noise from entities is worthwhile to be developed.
- This evaluation did not consider the use of an end-to-end model for the joint extraction of entities and relationships. A possible reason is that there are too many types of entities and relationships, but few train corpora. Therefore, developing an end-to-end model in this situation is challenging.
- Studies on extending the knowledge graph schema, such as 50 entity types, hundreds of entity attributes, and relationships between entities should be performed.
- Further research on automatic construction of multilingual knowledge graphs should be conducted. The evaluation in this paper did not take multilingualism into account, since our goal is a financial research report knowledge graph (FR2KG) in the Chinese language. In particular, the FR2KG data set did not involve the fusion of entities among multiple languages. Constructing knowledge graph from a multilingual corpus involves many new topics, including entity alignment and extraction relationships between different language entities. Simultaneously, it would be meaningful to evaluate the automatic construction of multilingual knowledge graphs.
- This evaluation implies the disambiguation and fusion of a small number of entities. There is no explicit evaluation of this area. In this regard, the evaluation of knowledge disambiguation and fusion will be more and more active in the future.
- It is a significant topic to study the difficulty of entity extraction, attribute extraction and relationship extraction in detail. In addition, it is also valuable and meaningful to set reasonable metrics for the automatic construction of knowledge graph.

## 7. CONCLUSIONS

In this paper, we introduce a high-quality data set, named financial research report knowledge graph (FR2KG), which consists of 17,799 entities, 26,798 relationship triples, and 1,328 attribute triples covering 10 entity types, 19 relationship types, and 6 attributes. We present an overview of the evaluation task of automated construction of Financial Knowledge Graph at CCKS2020. In addition, we summarized the technologies for automatically constructing knowledge graphs, and introduced some challenging topics and new directions for research in knowledge graphs.

## AUTHOR CONTRIBUTIONS

W.G. Wang (wangwenguang@datagrand.com) is the team leader for this project. He provided overall technical leadership, designed the data schema, performed curation work and contributed to writing and editing of the manuscript. Y.L. Xu (xuyonglin@datagrand.com) investigated the latest progress of Relation Extraction and wrote relevant chapters. C.H. Du (dunchunhui@datagrand.com) investigated the latest progress of Entity Extraction and wrote relevant chapters. Y.W. Chen (chenyunwen@datagrand.com) organized the data annotation and contributed to writing and editing of the manuscript as a senior author. Y.Y. Wang (wangyijie@datagrand.com) and H. Wen (wenhui@datagrand.com) contributed to the statistics and review of evaluation results. All the authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript.

## DATA AVAILABILITY STATEMENT

All the data are available at Data Intelligence's data repository at the Science Data Bank, <https://doi.org/10.11922/sciencedb.01060>, under an Attribution 4.0 International (CC BY 4.0).

## REFERENCES

- [1] Jia, D., et al.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
- [2] Ji, H., Nothman, J.: Overview of TAC-KBP2016 tri-lingual EDL and its impact on end-to-end KBP. In: Proceedings of Text Analysis Conference, pp. 1–15 (2016)
- [3] Elhammadi, S., et al.: A high precision pipeline for financial knowledge graph construction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 967–977 (2020)
- [4] Ein-Dor, L., et al.: Financial event extraction using Wikipedia-based weak supervision. In: Proceedings of the Second Workshop on Economics and Natural Language Processing, pp. 10–15 (2019)
- [5] TAC KBP 2016 Cold Start Track. Available at: <https://tac.nist.gov/2016/KBP/ColdStart/index.html>. Accessed 30 July 2021
- [6] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186 (2019)
- [7] Zhang, Z., et al.: ERNIE: Enhanced language representation with informative entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1441–1451 (2019)
- [8] Ringland, N., et al.: NNE: A data set for nested named entity recognition in English newswire. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5176–5181 (2019)
- [9] Thorne, J., et al. FEVER: A large-scale data set for fact extraction and VERification. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 809–819 (2018)
- [10] Yao, Y., et al: DocRED: A large-scale document-level relation extraction data set. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 764–777 (2019)
- [11] Zhu, T., et al.: Towards accurate and consistent evaluation: A data set for distantly-supervised relation extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6436–6447 (2020)

- [12] Wang, L., et al.: CORD-19: The Covid-19 open research data set. arXiv preprint arXiv:2004.10706v2 (2020)
- [13] D'Souza, J., et al.: The STEM-ECR data set: Grounding scientific entity references in STEM scholarly content to authoritative encyclopedic and lexicographic sources. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2192–2203 (2020)
- [14] Sang, E.F., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 142–147 (2003)
- [15] BOSON data set. Available at: <https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/boson>. Accessed 30 June 2021
- [16] People's Daily data set. Available at: <https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/renMinRiBao>. Accessed 30 June 2021
- [17] Levow, G.A.: The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 108–117 (2006)
- [18] Mikolov, T., et al.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations, pp. 1–12 (2013)
- [19] Yao, L., et al.: Biomedical named entity recognition based on deep neural network. International Journal Hybrid Information Technology 8(8), 279–288 (2015)
- [20] Nguyen, T.H., et al.: Toward mention detection robustness with recurrent neural networks. arXiv preprint arXiv:1602.07749 (2016)
- [21] Zheng, S., et al.: Joint extraction of entities and relations based on a novel tagging scheme. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1227–1236 (2017)
- [22] Huang, Z., et al.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
- [23] Li, P.H., et al.: Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2664–2669 (2017)
- [24] Wang, C., et al.: Code-switched named entity recognition with embedding attention. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching (CALCS), pp. 154–158 (2018)
- [25] Kuru, O., et al.: CharNER: Character-level named entity recognition. In: Proceedings of the 26th International Conference on Computational Linguistics, pp. 911–921 (2016)
- [26] Peters, M.E., et al.: Deep contextualized word representations. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 2227–2237 (2018)
- [27] Peters, M.E., et al.: Semi-supervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1756–1765 (2017)
- [28] Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
- [29] Liu, T., et al.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5301–5307 (2019)
- [30] Song, C.H., et al.: Improving neural named entity recognition with gazetteers. arXiv preprint arXiv:2003.03072 (2020)
- [31] Jie, Z., Lu, W.: Dependency-guided LSTM-CRF for named entity recognition. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3860–3870 (2019)

- [32] Liu, Y., et al.: GCDT: A global context enhanced deep transition architecture for sequence labeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2431–2441 (2019)
- [33] Luo, Y., et al.: Hierarchical contextualized representation for named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8441–8448 (2019)
- [34] Collobert, R., et al.: Natural language processing (Almost) from scratch. *Journal of Machine Learning Research* 12, 1462–1467 (2011)
- [35] Strubell, E., et al.: Fast and accurate entity recognition with Iterated Dilated Convolutions. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2670–2680 (2017)
- [36] Lample, G., et al.: Neural architectures for named entity recognition. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 260–270 (2016)
- [37] Chiu, J.P.C., et al.: Named entity recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4, 357–370 (2016)
- [38] Chaudhary, A., et al.: A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 5163–5173 (2019)
- [39] Zhai, F., et al.: Neural models for sequence chunking. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3365–3371 (2017)
- [40] Shen, Y., et al.: Deep active learning for named entity recognition. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 252–256 (2017)
- [41] Li, X., et al.: Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 465–476 (2020)
- [42] Liu, X., et al.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 359–367 (2011)
- [43] Etzioni, O., et al.: Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence* 165(1), 91–134 (2005)
- [44] Zhang, S., Elhadad, N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics* 46(6), 1088–1098 (2013)
- [45] Nadeau, D., et al.: Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In: Conference of the Canadian Society for Computational Studies of Intelligence, pp. 266–277 (2006)
- [46] Brooke, J., et al.: Bootstrapped text-level named entity recognition for literature. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 344–350 (2016)
- [47] Jia, C., et al.: Cross-domain NER using cross-domain language modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2464–2474 (2019)
- [48] Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 100–110 (1999)
- [49] Hendrickx, I., et al.: SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 33–38 (2010)
- [50] Li, S., et al.: DuIE: A large-scale Chinese data set for information extraction. In: The CCF International Conference on Natural Language Processing and Chinese Computing, pp. 791–800 (2019)
- [51] Zeng, D., et al.: Relation classification via Convolutional Deep Neural Network. In: Proceedings of the 25th International Conference on Computational Linguistics, pp. 2335–2344 (2014)

- [52] Santos, C.N., et al.: Classifying relations by ranking with Convolutional Neural Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 626–634 (2015)
- [53] Xu, K., et al.: Semantic relation classification via Convolutional Neural Networks with simple negative sampling. In: EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 536–540 (2015)
- [54] Xu, Y., et al.: Classifying relations via long short term memory networks along shortest dependency paths. In: EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 1785–1794 (2015)
- [55] Zhang, Y., et al.: Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2205–2215 (2018)
- [56] Guo, Z., et al.: Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 241–251 (2020)
- [57] Mintz, M., et al.: Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 1003–1011 (2009)
- [58] Zeng, D., et al.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1753–1762 (2015)
- [59] Jiang, X., et al.: Relation extraction with multi-instance multi-label convolutional neural networks. In: Proceedings of the 26th International Conference on Computational Linguistics, pp. 1471–1480 (2016)
- [60] Ji, G., et al.: Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, pp. 3060–3066 (2017)
- [61] Lin, Y., et al.: Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 2124–2233 (2016)
- [62] Bordes, A., et al.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, pp. 2787–2795 (2013)
- [63] Du, J., et al.: Multi-level structured self-attentions for distantly supervised relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2216–2225 (2018)
- [64] Vashishth, S., et al.: RESIDE: Improving distantly-supervised neural relation extraction using side information. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1257–1266 (2018)
- [65] Wang, G., et al.: Label-free distant supervision for relation extraction via knowledge graph embedding. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2246–2255 (2018)
- [66] Hasegawa, T., et al.: Discovering relations among named entities from large corpora. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 415–422 (2004)
- [67] Rozenfel, B., Ronen, F.: High-performance unsupervised relation extraction from large corpora. In: The Sixth International Conference on Data Mining, pp. 1032–1037 (2006)
- [68] Davidov, D., et al.: Fully unsupervised discovery of concept-specific relationships by Web mining. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 232–239 (2007)
- [69] Yan, Y., et al.: Unsupervised relation extraction by mining Wikipedia texts using information from the Web. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1021–1029 (2009)

- [70] Bollegala, D.T., et al.: Measuring the similarity between implicit semantic relations from the Web. In: Proceedings of the 18th International Conference on World Wide Web, pp. 651–660 (2009)
- [71] Miwa, M., et al.: End-to-end relation extraction using LSTMs on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1105–1116 (2016)
- [72] Zeng, X., et al.: Extracting relational facts by an end-to-end neural model with copy mechanism. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 506–514 (2018)
- [73] Fu, T.J., et al.: GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1409–1418 (2019)
- [74] Sun, C., et al.: Extracting entities and relations with joint minimum risk training. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2256–2265 (2018)
- [75] Takanobu, R., et al.: A hierarchical framework for relation extraction with reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7072–7079 (2019)
- [76] Li, X., et al.: Entity-relation extraction as multi-turn question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1340–1350 (2019)

## AUTHOR BIOGRAPHY



**Wenguang Wang** received his M.S. degree from Zhejiang University, China. He is currently Vice President of DataGrand Inc. His research interests include knowledge graph, natural language processing, computer vision, deep learning, reinforcement learning and content generation. He has ten patents and several academic publications in artificial intelligence. He is a member of China Computer Federation (CCF), Chinese Association for Artificial Intelligence (CAAI) and Chinese Information Processing Society of China (CIPS).

ORCID: 0000-0002-9617-0818



**Yonglin Xu** is currently an algorithm engineer at DataGrand Inc. He received his Master's degree from Shanghai University in 2019. His research interests include information extraction and knowledge graph.

ORCID: 0000-0002-7716-0841



**Chunhui Du** received his B.S. degree from the University of Electronic Science and Technology of China in 2018. He is currently pursuing a PhD degree from the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China. His research interests include federated learning and natural language processing.

ORCID: 0000-0002-9086-6021



**Yunwen Chen** received his PhD degree from Fudan University, China. He is the founder and the CEO of DataGrand Inc., Shanghai, a leading AI company in China. He was the Chief Data Officer of Shanda, Inc., Burlington, IA, USA, the senior Director of Tencent, Inc., Shenzhen, China, and a researcher of Baidu, Inc., Beijing, China. He has 32 patents and several academic publications. His current research interests include data mining, natural language processing, search and recommendation systems, and knowledge graphs. Dr. Chen was a recipient of the Distinguished Graduate Student in 2008. He is a Senior Member of the China Computer Federation (CCF) and a member of the ACM.

ORCID: 0000-0003-4513-9439



**Yijie Wang** is an algorithm engineer at DataGrand Inc. He received his Master's degree from Northeastern University in 2018. His research interests include data mining and knowledge graph.

ORCID: 0000-0003-1310-7467



**Hui Wen** graduated with a Master's degree in Computer Application Technology from Tongji University in 2014. He is a co-founder and senior scientist of DataGrand Inc, responsible for the research and development of knowledge graphs and data mining. He has rich R&D experience and strong interests in knowledge graphs, recommendation systems, search systems, distributed systems and so on.

ORCID: 0000-0001-7593-8544